GroupContrast: Semantic-aware Self-supervised Representation Learning for 3D Understanding

Chengyao Wang[1] Li Jiang[3] Xiaoyang Wu[2] Zhuotao Tian[4]
Bohao Peng[1] Hengshuang Zhao[2*] Jiaya Jia[1,4]

[1]CUHK [2]HKU [3]CUHK(SZ) [4]SmartMore

https://github.com/dvlab-research/GroupContrast

Abstract

Self-supervised 3D representation learning aims to learn effective representations from large-scale unlabeled point clouds. Most existing approaches adopt point discrimination as the pretext task, which assigns matched points in two distinct views as positive pairs and unmatched points as negative pairs. However, this approach often results in semantically identical points having dissimilar representations, leading to a high number of false negatives and introducing a "semantic conflict" problem. To address this issue, we propose GroupContrast, a novel approach that combines segment grouping and semantic-aware contrastive learning. Segment grouping partitions points into semantically meaningful regions, which enhances semantic coherence and provides semantic guidance for the subsequent contrastive representation learning. Semantic-aware contrastive learning augments the semantic information extracted from segment grouping and helps to alleviate the issue of "semantic conflict". We conducted extensive experiments on multiple 3D scene understanding tasks. The results demonstrate that GroupContrast learns semantically meaningful representations and achieves promising transfer learning performance.

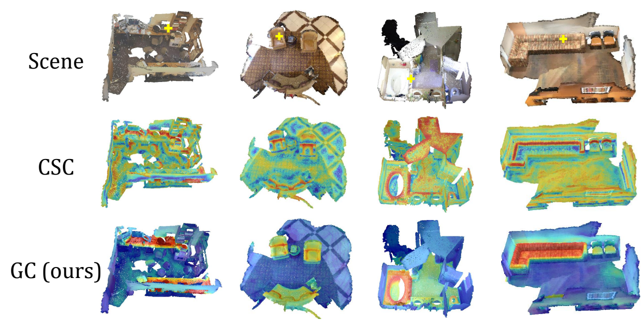

Figure 1. Visualization of activation maps depicting cosine similarity to the query point (indicated by a yellow cross) in the scene. Our approach demonstrates superior effectiveness in discriminating semantically similar points compared to CSC [20].

1. Introduction

Self-supervised visual representation learning aims to learn effective representations from large-scale unlabeled data. The learned representation can boost large amounts of downstream applications, such as object detection and semantic segmentation. Despite self-supervised learning has achieved remarkable results in 2D dense prediction tasks, 3D representation learning remains an emerging field. The predominant approaches for 3D scene recognition rely on supervised learning, where models are trained from scratch on specific datasets and tasks.

Recent studies [20, 43, 47] have explored the application of point discrimination as a pretext task for 3D self-supervised representation learning. These approaches transform each scene into two distinct views. They consider matched points between the two views as positive pairs and unmatched points as negative pairs. Despite decent performance gains observed in downstream tasks fine-tuning, there is a significant issue in previous point discrimination works: they focus on the geometric and structural relationships, disregarding the inherent semantic correlations. Consequently, they struggle to generate similar representations for semantically similar points within the 3D scene. To illustrate this issue, we visualize the activation map of a previous point discrimination model, CSC [20], as shown in Figure 1. The visualization reveals that previous methods fail to effectively capture semantic similarity. High-response points are scattered throughout the scene despite not necessarily possessing semantic similarity with query points. Conversely, points that are semantically similar to the query point may exhibit low correlations.

*Corresponding author.

This observation motivates us to improve the point discrimination pretext task. Merely designating unmatched points as negative pairs in the pretext task may result in a high number of false negatives. This is because elements that should be semantically identical are compelled to have dissimilar representations, which we refer to as the "*semantic conflict*". As a consequence, the conflict may compromise the semantic consistency that is crucial for downstream dense prediction tasks where different individuals are required to be assigned to their corresponding semantic labels. To address this issue, we present GroupContrast which consists of two essential parts: 1) Segment Grouping and 2) Semantic-Aware Contrastive Learning.

Segment grouping aims to enhance the semantic coherence among points within a scene and provide semantic guidance for the following contrastive learning. It achieves this by partitioning the point cloud into semantically similar groups via a segment-level deep clustering process. In particular, we first generate initial segments via a graph cut method based on low-level geometric information [12] and get the segment features via segment-wise pooling. We employ a set of learnable prototypes as cluster centers. Correlations between segment representations and these prototypes are then computed, and an informative-aware distillation loss is applied to encourage consistency between the segment-prototype correlations across two views with different augmentations. Segment grouping holds significant potential for effectively grouping semantically similar segments, thereby serving as a robust foundation for advancing point discrimination and addressing "*semantic conflict*".

Semantic-aware contrastive learning. Based on the results of segment grouping, we can improve the pretext task of point discrimination by integrating the positive pairs obtained within the same group and the negative pairs derived from different groups. This approach helps to alleviate the issue of "*semantic conflict*" by ensuring that the elements in negative pairs have distinct geometric representations in the representation space. An InfoNCE loss [25] is then applied to aggregate positive pairs and scatter negative pairs in the representation space. Besides, the confidence weight is found conducive to contrastive learning by mitigating the adverse impacts of incorrect segment assignments yielded by segment grouping.

As shown in the activation map depicted in Figure 1, our method effectively recognizes semantically similar points in the scene for the query point, in contrast to the confusion observed in the CSC model [20]. This highlights the emerging capacity of GroupContrast in semantic-level recognition. Extensive experiments on 3D semantic segmentation, instance segmentation, and object detection demonstrate the promising transfer learning performance of GroupContrast. For instance, our approach achieves 75.7% mIoU on ScanNet [10] and 30.0% mIoU on ScanNet200 [32] semantic segmentation using a SparseUNet [9] pre-trained by our method. These results outperform current state-of-the-art self-supervised 3D representation learning approaches. The contribution of our work can be summarized as follows:

- We examine the representations generated by the current unsupervised point cloud representation learning method and observe the presence of semantic conflict, which can potentially impede the performance of downstream applications.
- We propose GroupContrast, consisting of Segment Grouping and Semantic-aware Contrastive Learning, to address the semantic conflict by preserving the cross-view geometric consistency while avoiding negative pairs with similar semantics.
- Extensive experiments demonstrate that GroupContrast achieves state-of-the-art transfer learning results in various 3D scene perception tasks.

2. Related Work

2D self-supervised representation learning. Instance discrimination [11] as a pretext task for self-supervised visual representation learning has made remarkable progress in recent years. By leveraging InfoNCE loss [25] as an optimization objective for contrastive representation learning, a number of studies [5, 7, 17] have shown impressive transfer learning performance. More recently, modern approaches [2, 3, 6, 8, 15, 26] further improve this paradigm by removing negative pairs. Despite the impressive transfer learning performance on image classification tasks, instance discrimination treats an image as a whole, migrating complex structures in natural images. To address this issue, some studies explore pixel discrimination [23, 38, 48] and object discrimination [18, 19, 40, 41, 46] as a pretext task, which enhances the intrinsic structure of the image and further improves the transfer learning performance on dense prediction tasks. In this work, we attempt to conduct visual representation learning on complicated 3D scenes, which is more correlated with this line of work.

3D self-supervise representation learning. Unlike the 2D counterpart, self-supervised representation learning on 3D point clouds is still an emerging area. Earlier works [16, 33, 34, 39] conduct self-supervised representation learning on object-centric point clouds [4]. Experimentally, these approaches are unable to benefit 3D scene understanding [47]. Recent works [20, 21, 43, 47, 49, 51, 55] start to build self-supervised 3D representation learning on scene-centric data [10] and found sufficient performance improvement when transferred to a diverse set of 3D scene perception tasks. As a pioneer work, PointContrast [47] adopts point discrimination as a pretext task. CSC [20] explores point discrimination with scene context descriptors. MSC [43]

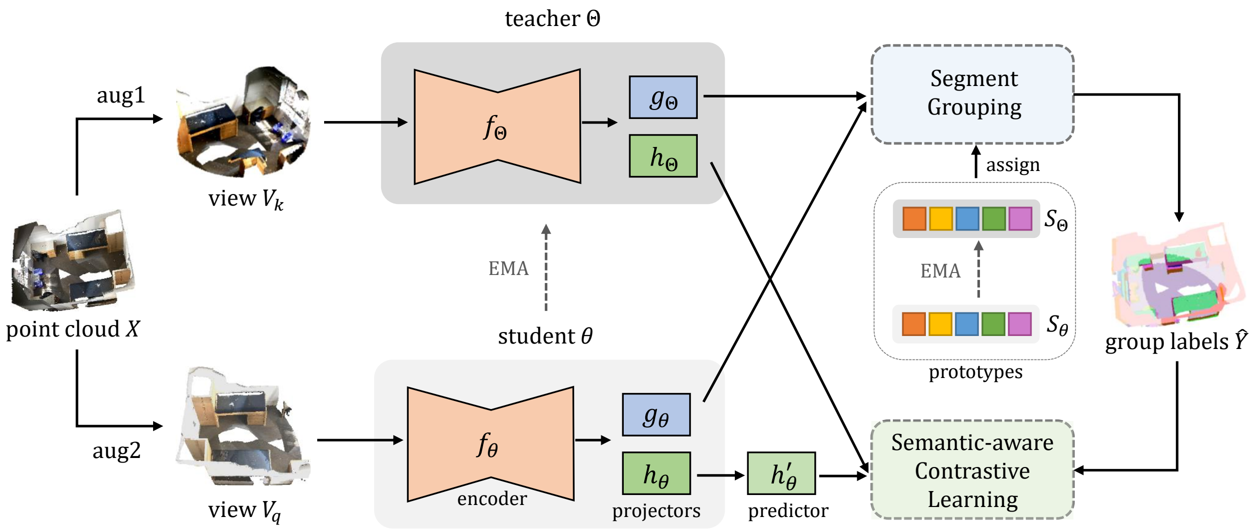

Figure 2. **Overview of our proposed GroupContrast framework.** Our framework uses two neural networks, each comprising a backbone and two projectors for segment grouping and contrastive learning. The parameters of the teacher network are updated as an exponential moving average (EMA) of the parameters of the student network. The student network includes an additional asymmetric predictor for contrastive learning. The Segment Grouping module assigns each point to one of n prototypes, and this clustering result serves as a guide for effective contrastive representation learning.

introduces masked reconstruction learning to enforce the pretext and alleviate the mode collapse problem. However, these approaches treat matched points as positive pairs and unmatched points as negative pairs, leading to a large number of false negatives. In contrast, we attempt to discover semantic meaningful regions to avoid the model being confused by false negatives. Cluster3Dseg [13] group points with identical labels into subclass and learn a better representation space via contrastive learning. But they focus on supervised learning, while we focus on unsupervised representation learning.

3D scene understanding. There are two primary architectures for 3D scene understanding: point-based and voxel-based methods. Point-based methods [29, 30, 42, 44, 50, 52–54] directly operate on the points, making them well-suited for learning point clouds. However, directly operating on the points makes this line of work computationally expensive. In contrast, voxel-based methods transform points into regular voxels to apply 3D convolutions [24, 27, 36]. Driven by highly optimized sparse convolution [9, 14], this line of work achieves excellent efficiency. Following previous works on 3D representation learning [43, 45, 47], we conduct representation learning and downstream task fine-tuning on a voxel-based method SparseUNet [9].

3. Method

In this section, we introduce a novel method, GroupContrast, for 3D self-supervised representation learning to enhance the feature alignment among semantically similar points. GroupContrast consists of two key components: Segment Grouping and Semantic-aware Contrastive Learning. Firstly, we present the overall framework of our GroupContrast in Section 3.1. Then, we delve into Segment Grouping in Section 3.2, which enables the discovery of semantic meaningful regions in unlabeled point clouds. Following that, we introduce Semantic-aware Contrastive Learning in Section 3.3, which leverages the regions discovered in Segment Grouping for effective contrastive representation learning.

3.1. Overall Framework

In GroupContrast, we employ a dual-network structure, including a teacher network and a student network, to ensure a stable and consistent contrastive learning process. As illustrated in Figure 2, the student network θ consists of an encoder f_θ, two projectors g_θ and h_θ, an asymmetric predictor h'_θ, and a set of n learnable prototypes $S_\theta \in \mathcal{R}^{n\times D}$, where D indicates the feature dimension. The teacher network Θ shares the same architecture as the student network, except for the asymmetric prediction layer h'_θ. The teacher network has a different set of parameters, which are formed by taking the exponential moving average (EMA) of the student network θ.

Given a point cloud X, two augmented views V_k and V_q derived from X are fed into the teacher network and the student network, respectively. Then, we take the point-level features produced by the projectors g and h as the in-

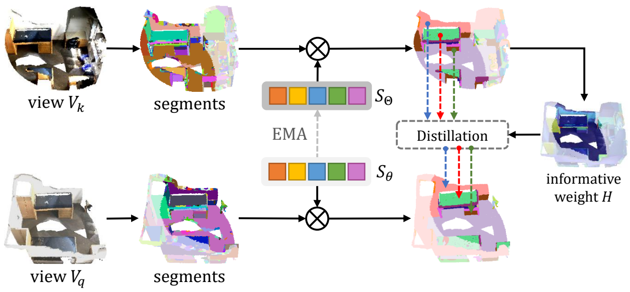

Figure 3. **Segment Grouping** is optimized by distilling the assignment scores between each segment and the n prototypes from the teacher network to the student network. An informative weight is employed to make the student network focus on more challenging segments.

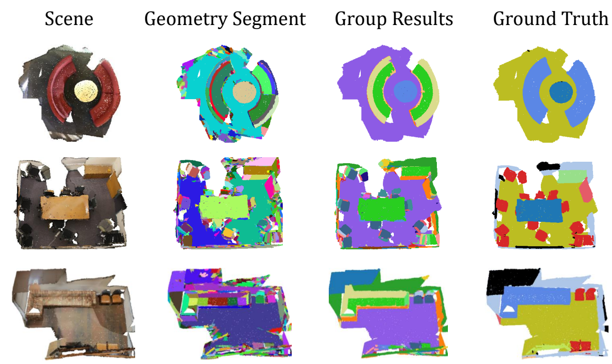

Figure 4. **The result of Segment Grouping.** We compare the grouping results with original geometry segments [12] and semantic ground truth. Segment grouping effectively groups points into semantically meaningful regions without human supervision.

puts for the Segment Grouping module and Semantic-aware Contrastive Learning module, respectively. In the Segment Grouping process, a set of learnable prototypes S are used as cluster centers for identifying the meaningful semantic groups within the 3D scene. These groups are then employed in the Semantic-aware Contrastive Learning module to mitigate the semantic conflict problem and assist the representation learning.

3.2. Segment Grouping

The Segment Grouping module is illustrated in Figure 3. We first utilize geometric information of the points (*e.g.*, normal) to generate P segments for the overlapped region of the two augmented views V_q and V_k using a graph cut method [12]. Then, segment-wise average pooling is applied on the l2-normalized point-level features produced by the projector g for both augmented views, resulting in the segment-level features $z_q \in \mathcal{R}^{P \times D}$ and $z_k \in \mathcal{R}^{P \times D}$. After that, we calculate the prototype assignment scores for each segment by measuring the cosine similarity between the segment-level features and the n learnable prototypes. Specifically, with segment-level features z_q, z_k and n learnable prototypes S_θ, S_Θ which are all l2-normalized, the assignment scores $Q \in \mathcal{R}^{P \times n}$ and $K \in \mathcal{R}^{P \times n}$ for segments in each view can be written as

$$Q = \underset{n}{\text{softmax}}(z_q S_\theta^{\mathrm{T}}/\tau_s), \quad K = \underset{n}{\text{softmax}}(z_k (S_\Theta - c)^{\mathrm{T}}/\tau_t). \tag{1}$$

Here, the temperature parameters τ_t and τ_s control the sharpness of the output distribution for the teacher network and student network, respectively. Additionally, a bias term c is introduced to avoid collapse, which will be further discussed later.

The teacher network is an average of consecutive student networks. Averaging model weights over training steps tends to produce a more accurate model [28, 37]. We can take advantage of this to set the optimization objective of segment grouping. A cross-entropy loss is applied to encourage the assignment scores of the student network consistent with the teacher network. The overall grouping loss is computed as the average cross-entropy loss across P segments:

$$\mathcal{L}^{group} = -\frac{1}{P} \sum_{i \in [0,P)} K_i \cdot log(Q_i). \tag{2}$$

Prevention of collapse. Directly applying this optimization objective will lead to collapse [3]. Inspired by DINO [3], we apply centring and sharpening for the momentum teacher outputs to avoid model collapse. For sharpening, we make the teacher temperature τ_t lower than student temperature τ_s to produce a sharper target to avoid uniform assignments. For centring, we use a bias term c to the teacher and reduce it from the prototypes when producing the teacher assignments. The bias term c is formed by taking the EMA of the output produced by the teacher network:

$$c = \lambda_c \cdot c + (1 - \lambda_c) \cdot \frac{1}{P} \sum_{i \in [0,P)} z_k\,[i]\, S_\Theta^{\mathrm{T}}, \tag{3}$$

where P stands for the number of segments, and λ_c refers to the momentum value. Intuitively, centring prevents one prototype from dominating the prototype assignment process.

Informative-aware distillation. The approach described above can be regarded as a knowledge distillation procedure from teacher network Θ to student network θ. However, treating all segments equally in distillation can lead to the model overlooking more informative segments. These segments are typically more difficult to assign prototypes, *i.e.*, with higher entropy, and should be paid with extra attention during distillation. Therefore, we use the entropy of teacher assignment scores K to measure each segment's

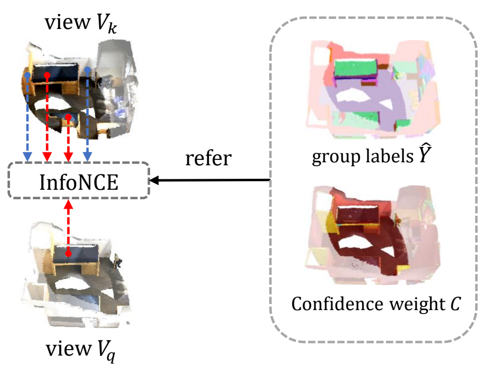

Figure 5. **Contrastive Learning.** We use an InfoNCE loss [25] to aggregate points within the same group and scatter points across different groups, as indicated by the Segment Grouping result. Here the red point in view V_q serves as a query, the red points in view V_k are positive samples, and the blue points in view V_k are negative samples. Both modules are conducted on overlapped regions of the two augmented views only, which are highlighted with darker colors in the figure.

"amount of information" and incorporate the entropy mask to distillation loss. The entropy mask H for each segment i can be concluded as $H_i = -\sum_{j=1}^{n} K_i^j \cdot log(K_i^j)$, and the grouping loss in Eq. 2 can be updated as

$$\mathcal{L}^{group} = -\frac{1}{\sum_{i\in[0,P)} H_i} \sum_{i\in[0,P)} H_i \cdot K_i \cdot log(Q_i). \tag{4}$$

Grouping result. With the assignment scores, we group the P segments into n clusters by assigning each segment to the prototype with the highest assignment score. We use the teacher assignment scores K to extract grouping results. Formally, the segment-level grouping result $\hat{Y}_{seg} \in \mathcal{R}^P$ is calculated as

$$\hat{Y}_{seg} = \operatorname*{argmax}_{n} K. \tag{5}$$

We then project $\hat{Y}_{seg}$ to each point to obtain the point-level group labels $\hat{Y}$ for the overlapped region of two views. As illustrated in Figure 4, our segment grouping process effectively groups initial segments into semantically meaningful regions.

3.3. Semantic-aware Contrastive Learning

As discussed in Section 1, the issue of "*semantic conflict*" exists in previous contrastive-based representation learning methods where the semantically identical elements may erroneously have distinct representations. To address this issue for achieving a better agreement between the semantically similar points, we use the group labels $\hat{Y}$ as the semantic guidance to enhance contrastive representation learning.

Semantic-aware positive pairs. As illustrated in Figure 5, we define the positive pairs for contrastive learning based on the semantic grouping result $\hat{Y}$. Points in the same group are set as positive pairs, while points in different groups are treated as negative pairs. Formally, for the two augmented views V_q and V_k, we sample N points from their overlapped region and set the point indices of these sampled points in $\hat{Y}$ as $I_q \in \mathcal{R}^N$ and $I_k \in \mathcal{R}^N$, respectively. The positive pair set is then defined as

$$\mathcal{P} = \{(i,j) | i \in I_q, j \in I_k, \hat{Y}_i = \hat{Y}_j\}. \tag{6}$$

Confidence-aware learning. In the early stages of training, the group labels $\hat{Y}$ may not always be reliable. Using noisy group labels for contrastive representation learning may confuse the model. Therefore, we evaluate the confidence of each positive pair and incorporate confidence weights in contrastive representation learning, to alleviate the adverse effects brought by the uncertain elements. Concretely, we leverage the teacher assignment scores K in confidence evaluation. For each positive pair i, j with grouping label k, the confidence weight $C_{i,j}$ is calculated as

$$C_{i,j} = K_{s_i,k} \times K_{s_j,k}, \tag{7}$$

where s_i and s_j indicate the segment indices of points i and j, respectively.

Improved contrastive loss. For the teacher network, we add a projector h_Θ after the encoder to extract feature v_k. For the student network, inspired by previous approaches [3, 5, 15], a projector h_θ together with an extra asymmetric predictor h'_θ is applied after encoder to extract feature v_q. Both v_q and v_k are l2-normalized for contrastive representation learning. The InfoNCE loss [25] is adopted to aggregate positive pairs and scatter negative pairs in the representation space. By incorporating the aforementioned confidence weight $C_{i,j}$, given a set of positive pairs $\mathcal{P}$ and a temperature parameter τ, the improved contrastive loss can be written as

$$\mathcal{L}^{con} = \frac{1}{|\mathcal{P}|} \sum_{i,j\in\mathcal{P}} -C_{i,j} \cdot \log \frac{\exp(v_q^i \cdot v_k^j / \tau)}{\exp(v_q^i \cdot v_k^j / \tau) + \sum\limits_{i,k\notin P} \exp(v_q^i \cdot v_k^k / \tau)}. \tag{8}$$

We set τ to 0.4, following previous approaches [20, 47].

3.4. Overall Optimization Objective

We jointly optimize segment grouping and contrastive representation learning for pre-training. The overall optimization objective is a weighted sum of Eq. 4 and Eq. 8, which can be written as

$$\mathcal{L}^{overall} = \lambda_g \mathcal{L}^{group} + \lambda_c \mathcal{L}^{con}, \tag{9}$$

where λ_g and λ_c are scale factors. We empirically set $\lambda_g = \lambda_c = 1$, as our experiments suggest that the performance is robust to different scale factors.

4. Experiments

We conduct extensive experiments to validate the effectiveness of our proposed GroupContrast framework. We first perform ablation studies in Section 4.1 to demonstrate the efficacy of each proposed component, then compare our approach with previous state-of-the-art self-supervised 3D representation learning approaches in Section 4.2.

4.1. Main Properties

To assess the effectiveness and analyze the key properties of our GroupContrast, we conduct ablation experiments on its core design choices. As a default setting, we first self-supervised pre-train a SparseUNet [9] on ScanNet [10] dataset for 600 epochs. We then utilize ScanNet semantic segmentation as the downstream task and evaluate the performance using the mIoU (%) metric. The results of ablation studies are concluded in Table 1. Please refer to the supplementary material for more implementation details about the pre-training and fine-tuning.

Positive pair construction. In Table 1a, we compare the semantic-aware positive pairs generated based on Segment Grouping results with several baselines to validate the effectiveness of Segment Grouping on contrastive representation learning. The baseline approaches include (1) Matched Points, which uses matched points in two views as positive pairs and unmatched points as negative pairs, similar to PointContrast[47]; (2) Spatial Grid, which divides the point cloud into multiple spatial grids and assigns points within the same grid as positive pairs. In this case, we set the grid size to $1m \times 1m$; and (3) Geometry Segment, which generates segments using a normal-based graph cut method [12] and assigns points within the same segment as positive pairs. As shown, simply assigning points in the same spatial grid as positive pairs leads to a decrease in transfer learning performance, even worse than the Matched Points baseline. Although Geometry Segments incorporate geometric priors into the network, the improvement in transfer learning is only marginal. By employing Segment Grouping based on Geometry Segments, we observe a noteworthy improvement of 1.1 points in transfer learning performance compared to the Matched Points baseline, affirming the effectiveness of our proposed segment grouping approach.

Number of prototypes. In Table 1b, we investigate the impact of the number of prototypes n in Segment Grouping. Our observations indicate that $n = 32$ yields the best performance for ScanNet pre-training. Too few prototypes can lead to excessive feature aggregation, while too many can cause the model to learn overly fine-grained features.

Number of sampled points. In Table 1c, we study the number of points sampled from the overlapped region for contrastive representation learning. We observe that the model is robust to the number of sampled points. For training efficiency, we sample 2048 points for each 3D scene.

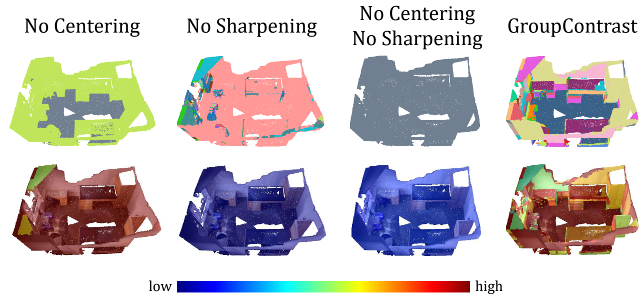

Figure 6. **Group labels** (top) and **Confidence weight** (down) generated from the pre-trained model without centring and/or sharpening. Without centring, points are grouped into one or two regions. Without sharpening, the assignment scores become uniform vectors, leading to low confidence weight.

Informative-aware distillation. In Table 1d, we study the effect of informative-aware distillation. Introducing informative weight prevents the model from overlooking more informative segments during distillation, leading to better transfer learning performance.

Teacher temperature τ_t. In Table 1e, we study the temperature parameter in Segment Grouping. We follow previous work [3] to set student temperature $\tau_s = 0.1$ and ablate different values of teacher temperature τ_t. The results show that a softer teacher leads to better transfer learning performance.

Predictor. In Table 1f, we study the effect of the asymmetric predictor. Similar to previous 2D self-supervised representation learning approaches [3, 5, 15, 41], introducing an asymmetric predictor makes the contrastive objective more challenging, resulting in better transfer learning performance.

Semantic-aware contrastive representation learning. In Table 1g, we study the design of semantic-aware contrastive representation learning. Introducing semantic-aware positive pairs for contrastive learning helps alleviate the issue of "*semantic conflict*" and results in better transfer learning performance. Moreover, incorporating confidence weight for each positive pair alleviates the adverse effects brought by the uncertain elements, further improving the transfer learning performance.

Collapse problem. In Table 1h, we study the effectiveness of centring and sharpening in avoiding collapse. As shown in the table, both centring and sharpening effectively improve the transfer learning performance on semantic segmentation. Furthermore, to further verify the effectiveness of centering and sharpening, we visualize the grouping results and corresponding assignment scores for GroupContrast without sharpening and centring in Figure 6. Without

Positive Pairs	FT mIoU(%)
Matched Points	74.6
Spatial Grid	74.2
Geometry Segment	74.8
Segment Grouping	**75.7**

(a) **Positive Pairs.** Positive pairs constructed based on Segment Grouping work best in downstream transfer.

Prototypes n	FT mIoU(%)
16	75.2
32	**75.7**
64	75.3
128	74.8

(b) **Number of prototypes.** 32 prototypes work best for Segment Grouping.

Sample Points	FT mIoU(%)
1024	75.5
2048	**75.7**
4096	75.5
8192	75.4

(c) **Number of sampled points.** Our approach is robust to number of sampled points.

Infomative-aware	FT mIoU(%)
	75.4
✓	**75.7**

(d) **Informative-aware.** Incorporating informative weight for distillation boosts downstream performance.

Temperature τ_t	FT mIoU(%)
0.04	75.3
0.07	**75.7**

(e) **Teacher temperature.** A softer teacher leads to better downstream performance.

Predictor	FT mIoU(%)
	75.3
✓	**75.7**

(f) **Predictor.** Including asymmetric predictor boosts downstream performance.

Semantic-aware	Confidence-aware	FT mIoU(%)
		74.8
✓		75.1
✓	✓	**75.7**

(g) **Semantic-aware Contrastive Learning.** Incorporating semantic-aware positive pairs and confidence weights can improve downstream performance.

Centering	Sharpening	FT mIoU(%)
		72.7
✓		73.1
	✓	74.1
✓	✓	75.7

(h) **Avoid Collapse.** Incorporating both centering and sharpening helps our approach to alleviating the collapse problem.

Epochs	ScanNet	ScanNet (20%)
300	74.8	65.0
600	**75.7**	65.8
1200	**75.7**	**66.5**

(i) **Pre-training epochs.** Scaling up the number of pre-train epochs makes the model more data-efficient.

Table 1. **Ablation Study.** Without further explanation, we pre-train a SparseUNet [9] for 600 epochs on ScanNet [10] dataset to analyse our main design choices and properties. We report fine-tuning mIoU results(%) on ScanNet Semantic Segmentation. Default settings are marked in gray.

centring, all points in a scene are grouped into two regions, resulting in multiple false positives for contrastive representation learning. Without sharpening, the assignment scores Q and K become uniform vectors, resulting in low confidence scores and noisy group labels, which make it hard for contrastive loss to converge. The result is even worse without both techniques, where all points are grouped into an identical region, and the corresponding confidence scores for each point are also low.

Pre-training epochs. In Table 1i, we study the number of pre-training epochs. Results indicate that increasing the pre-training epoch from 600 to 1200 does not enhance performance when fine-tuning the full ScanNet training set for semantic segmentation. However, in a data-efficient scenario where only 20% of reconstructed point clouds are used for fine-tuning, scaling up the pre-training epoch effectively improves performance.

4.2. Results Comparison

In this section, we evaluate the effectiveness of our proposed GroupContrast by comparing it with previous self-supervised 3D representation learning approaches [20, 43, 47]. Experiments are conducted on various downstream tasks, including 3D semantic segmentation, instance segmentation, and object detection. Additionally, we evaluate the data efficiency of GroupContrast on data-efficient 3D semantic segmentation. We apply SparseUNet [9] as the backbone and adopt a longer training schedule (1200 epochs). The transfer learning results are concluded in Table 2. Please refer to supplementary materials for more implementation details about downstream task fine-tuning.

Semantic segmentation. In Table 2a, we report the Semantic Segmentation results on ScanNet [10], ScanNet200 [32] and S3DIS [1] benchmark with a SparseUNet backbone. For in-domain transfer learning where pre-training and fine-tuning are both conducted on the ScanNet [10] dataset, we achieve 75.7% mIoU on the ScanNet validation set and 30.0% mIoU on the ScanNet200 validation set, outperforming the current state-of-the-art approaches by 0.7% mIoU and 1.2% mIoU, respectively. Moreover, the model pre-trained with our approach on ScanNet achieves 72.0% mIoU when transferred to S3DIS semantic segmentation, demonstrating that GroupContrast is also effective for cross-domain transfer learning.

Instance segmentation. In Table 2b, we report the instance segmentation results on ScanNet [10], ScanNet200 [32] and S3DIS [1] benchmark with PointGroup [22] as the instance segmentation head and SparseUNet as the backbone. We still find consistent transfer learning performance improvement compared with previous results. Specifically, our approach achieves 62.3 mAP@0.5 on the ScanNet validation set, which is 2.7 points higher than the previous state-of-the-art 3D self-supervised pre-training method. Furthermore, our approach achieves 27.5 mAP@0.5 when fine-tuning

Datasets	Semantic Segmentation (mIoU)				
	SC	PC[47]	CSC[20]	MSC[43]	GC(ours)
ScanNet	72.2	74.1	73.8	75.3	**75.7**
ScanNet200	25.0	26.2	26.4	28.8	**30.0**
S3DIS	68.2	70.3	**72.2**	-	72.0

(a) **Semantic Segmentation.** We report the mIoU (%) results on ScanNet, ScanNet200, and S3DIS benchmarks.

Datasets	Instance Segmentation (mAP@0.5)				
	SC	PC[47]	CSC[20]	MSC[43]	GC(ours)
ScanNet	56.9	58.0	59.4	59.6	**62.3**
ScanNet200	24.5	24.9	25.2	26.8	**27.5**
S3DIS	59.3	60.5	63.4	-	**63.5**

(b) **Instance Segmentation.** We report the mAP@0.5 results on ScanNet, ScanNet200, and S3DIS benchmarks.

Datasets	Object Detection (mAP@0.5)				
	SC	PC[47]	CSC[20]	MSC[43]	GC(ours)
ScanNet	35.2	38.0	39.3	-	**41.1**
SUN RGB-D	31.7	34.8	36.4	-	**37.0**

(c) **Object Detection.** We report the mAP@0.5 results on ScanNet and SUN RGB-D benchmarks.

Table 2. **Results comparison** on 3D Semantic Segmentation, Instance Segmentation and Object Detection. We pre-train our approach on ScanNet point cloud with SparseUNet[9] as the backbone for transfer learning performance comparison. *SC* denotes train from scratch. Our results are marked in gray.

on ScanNet200 instance segmentation and 63.5 mAP@0.5 when fine-tuning on S3DIS instance segmentation, both outperforming previous results.

Object detection. In Table 2c, we report the Object Detection results on ScanNet [10] and SUN-RGBD [35] benchmark with VoteNet [31] as the detection head and SparseUNet as the backbone. As shown in the table, we can also find performance improvement in object detection fine-tuning. We can achieve 41.1 mAP@0.5 on the ScanNet validation set and 37.0 mAP@0.5 on the SUN-RGBD validation set, surpassing the previous state-of-the-art 3D self-supervised pre-training approach by 1.8 points and 0.6 points respectively.

Data efficiency. Apart from full dataset fine-tuning, we also evaluate the data efficiency of our approach on the ScanNet Data Efficient Semantic Segmentation benchmark. We use the same data split as CSC [20] in both limited scene reconstruction and limited points annotation settings. We report the data efficient semantic segmentation result in Table 3. As reported, our approach achieves state-of-the-art performance for all cases in both settings. These results suggest that our proposed approach is effective in improving the data efficiency of 3D scene understanding.

LR	Semantic Segmentation (mIoU)			
Pct.	SC	CSC [20]	MSC [43]	GC (ours)
100%	72.2	73.8	75.0	**75.7**
1%	26.1	28.9	29.2	**30.7**
5%	47.8	49.8	50.7	**52.9**
10%	56.7	59.4	61.0	**62.0**
20%	62.9	64.6	64.9	**66.5**

(a) **Limited Reconstruction.** We compare the mIoU (%) results on ScanNet data efficient semantic segmentation benchmark with limited scene reconstruction setting.

LA	Semantic Segmentation (mIoU)			
Pct.	SC	CSC [20]	MSC [43]	GC (ours)
Full	72.2	73.8	75.0	**75.7**
20	41.9	55.5	**61.2**	**61.2**
50	53.9	60.5	66.8	**67.3**
100	62.2	65.9	69.7	**70.3**
200	65.5	68.2	70.7	**71.8**

(b) **Limited Annotation.** We compare the mIoU (%) results on ScanNet data efficient semantic segmentation benchmark with limited point annotation setting.

Table 3. **Data Efficiency.** We evaluate the data efficiency of GroupContrast on ScanNet data efficient semantic segmentation benchmark. The model is pre-trained on ScanNet point cloud with SparseUNet[9] as the backbone. *SC* denotes train from scratch. Our results are marked in gray.

5. Conclusion

This work presents GroupContrast, a self-supervised representation learning framework for 3D understanding, with joint segment grouping and semantic-aware contrastive learning. Segment grouping discovers semantically meaningful regions by assigning each segment a prototype. Based on the grouping result, a contrastive learning objective is applied to produce a semantic-aware representation space. Our approach can effectively decompose a point cloud into multiple semantically meaningful regions without supervision, showing the emerging ability in semantic-level recognition. Moreover, extensive experimental results demonstrate that our approach achieves promising transfer learning performance on 3D semantic segmentation, object detection and instance segmentation.

While our approach brings great benefits to downstream tasks through ScanNet pre-training, it is currently limited by the relatively small scale of the pre-training data. As a direction for future research, we aim to explore cross-dataset pre-training to enlarge the data size and collaborate our framework with well-trained foundation models. These efforts are expected to overcome this limitation and enhance the generalizability and robustness of our framework.

Acknowledgements. This work receives partial support from the Shenzhen Science and Technology Program under No. KQTD20210811090149095.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 7

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 5, 6

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 5, 6

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021. 2

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2, 3, 6, 7, 8

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6, 7, 8

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2, 4, 6

[13] Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, and Qinghua Zheng. Clustering based point cloud representation learning for 3d analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8283–8294, 2023. 3

[14] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 3

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 5, 6

[16] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *ICCV*, 2019. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[18] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 2

[19] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 123–143. Springer, 2022. 2

[20] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1, 2, 5, 7, 8

[21] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2

[22] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *CVPR*, 2020. 7

[23] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2

[24] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015. 3

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[27] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Omni-adaptive sparse cnns for 3d semantic segmentation. 3

[28] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 4

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 3

[30] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 3

[31] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 8

[32] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 2, 7

[33] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *ECCV*, 2020. 2

[34] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *NeurIPS*, 2019. 2

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 8

[36] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 3

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 4

[38] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[39] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, 2019. 2

[40] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 2

[41] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, 2022. 2, 6

[42] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 3

[43] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, 2023. 1, 2, 3, 7, 8

[44] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 3

[45] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *CVPR*, 2024. 3

[46] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021. 2

[47] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[48] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 2

[49] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Unipad: A universal pre-training paradigm for autonomous driving. In *CVPR*, 2024. 2

[50] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 3

[51] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[52] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, 2019. 3

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

[54] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. 2023. 3

[55] Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, and Wanli Ouyang. Ponderv2: Pave the way for 3d foundation model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023. 2

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 7

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 5, 6

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 5, 6

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021. 2

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2, 3, 6, 7, 8

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6, 7, 8

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2, 4, 6

[13] Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, and Qinghua Zheng. Clustering based point cloud representation learning for 3d analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8283–8294, 2023. 3

[14] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 3

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 5, 6

[16] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *ICCV*, 2019. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[18] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 2

[19] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 123–143. Springer, 2022. 2

[20] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1, 2, 5, 7, 8

[21] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2

[22] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *CVPR*, 2020. 7

[23] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2

[24] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015. 3

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[27] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Omni-adaptive sparse cnns for 3d semantic segmentation. 3

[28] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 4

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 3

[30] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 3

[31] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 8

[32] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 2, 7

[33] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *ECCV*, 2020. 2

[34] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *NeurIPS*, 2019. 2

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 8

[36] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 3

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 4

[38] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[39] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, 2019. 2

[40] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 2

[41] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, 2022. 2, 6

[42] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 3

[43] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, 2023. 1, 2, 3, 7, 8

[44] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 3

[45] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *CVPR*, 2024. 3

[46] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021. 2

[47] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[48] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 2

[49] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Unipad: A universal pre-training paradigm for autonomous driving. In *CVPR*, 2024. 2

[50] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 3

[51] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[52] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, 2019. 3

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

[54] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. 2023. 3

[55] Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, and Wanli Ouyang. Ponderv2: Pave the way for 3d foundation model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023. 2